Published in final edited form as:
Biometrics. 2010 March ; 66(1): 97–104. doi:10.1111/j.1541-0420.2009.01274.x.

Variable Selection in the Cox Regression Model with Covariates Missing at Random

Ramon I. Garcia, **Joseph G. Ibrahim***, and **Hongtu Zhu**
Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599-7420, U.S.A

Summary

We consider variable selection in the Cox regression model (Cox, 1975, *Biometrika* **362,** 269–276) with covariates missing at random. We investigate the smoothly clipped absolute deviation penalty and adaptive least absolute shrinkage and selection operator (LASSO) penalty, and propose a unified model selection and estimation procedure. A computationally attractive algorithm is developed, which simultaneously optimizes the penalized likelihood function and penalty parameters. We also optimize a model selection criterion, called the IC_Q statistic (Ibrahim, Zhu, and Tang, 2008, *Journal of the American Statistical Association* **103,** 1648–1658), to estimate the penalty parameters and show that it consistently selects all important covariates. Simulations are performed to evaluate the finite sample performance of the penalty estimates. Also, two lung cancer data sets are analyzed to demonstrate the proposed methodology.

Keywords

ALASSO; Missing data; Partial likelihood; Penalized likelihood; Proportional hazards model; SCAD; Variable selection

1. Introduction

In the analysis of regression models for censored survival data, one primary objective is to assess the importance of certain prognostic factors such as age, gender, or race in predicting survival outcome. This objective is further complicated by the presence of missing covariates. This is a general problem that is encountered in most clinical trials research in cancer and AIDS. There is a vast literature on parameter estimation in the Cox regression model in the presence of missing covariates, including Schluchter and Jackson (1989); Lipsitz and Ibrahim (2000); Paik and Tsai (1997); Chen and Little (1999); Herring and Ibrahim (2001); and Chen (2002). When trying to perform variable selection in this scenario, it is common to use some model selection criteria, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), or deviance information criterion (DIC) (Celeux et al., 2006; Pettitt et al., 2006), to select a small set of "covariates" that best predicts the outcome of interest. In the presence of missing covariate data, this approach requires the calculation of the observed data likelihood, which is not available in a closed form and is very difficult to approximate accurately. Because these likelihood calculations

*ibrahim@bios.unc.edu.

6. Supplementary Materials

Web Appendix A referenced in Sections 2 and 3 is available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

are necessary for each of the models under consideration, model selection criteria based approaches can become infeasible for variable selection (Fan and Li, 2001, 2002). Alternatively, penalized likelihood methods (Fan and Li, 2001), which perform variable selection and estimation simultaneously, do not require these likelihood calculations for each of the models under consideration.

Penalized likelihood methods using the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and adaptive Lasso (ALASSO) (Zou, 2006) penalties have been successfully applied to various parametric and semiparametric models including Cox's proportional hazards model without the presence of missing covariates (Fan and Li, 2002; Zhang and Lu, 2007). Extending penalized likelihood methods to perform variable selection in the Cox model with missing covariates raises many new statistical challenges, underscoring the need for methodological development. The first challenge is in maximizing the observed data log-likelihood function along with the SCAD or ALASSO penalties to select important variables and calculate their estimates. As already noted, the observed data log likelihood for missing data problems is often not available in closed form, and is computationally intractable and infeasible because it involves complicated high dimensional integrals, and the accuracy of the approximation to such integrals is essentially impossible to verify in many cases. The second challenge is in selecting appropriate penalty parameters to produce efficient estimates with suitable asymptotic properties such as sparsity and asymptotic normality (Fan and Li, 2001). The primary method of selecting penalty parameters for survival models is to use the penalty parameter value which optimizes the generalized crossvalidation (GCV) (Fan and Li, 2002) criterion. It has been shown that for the linear model, the GCV cannot identify the true model consistently whereas the BIC can (Wang and Leng, 2007; Wang, Li, and Tsai, 2007). We expect that this is also the case for general statistical models including the Cox regression model. Also, this criterion needs to be well defined (Celeux et al., 2006) in the presence of missing data and should incorporate the parameters of the covariate distribution, which will need to be specified due to the presence of missing covariate data. To the best of our knowledge, a well-defined criterion and easy-to-compute penalty estimate are not currently available for the Cox regression model with missing covariate data.

The aim of this article is to develop a variable selection procedure and a consistent penalty estimation criterion based on the SCAD and ALASSO penalties for the Cox regression model with missing at random (MAR) covariates. We reformulate the penalty parameters in the SCAD and ALASSO penalty functions as hyperparameters of the regression coefficients. Then, we use the expectation–maximization (EM) algorithm to simultaneously optimize the penalized likelihood function and estimate the penalty parameters. In addition, we also develop an alternative method based on the IC_Q criterion to select penalty parameters. Under some regularity conditions, we establish the asymptotic properties of the maximum penalized likelihood (MPL) estimator and consistency of the IC_Q -based penalty estimation method.

To illustrate the proposed methodology, we consider data from a phase III advanced nonsmall-cell lung cancer (SCLC) clinical trial, labeled LCCC 9719, conducted by the University of North Carolina at Chapel Hill (Socinski et al., 2002). The goal of this trial is to compare a defined duration of therapy (A) to continuous therapy followed by a second line therapy (B) to determine optimal duration of therapy in SCLC patients. LCCC 9719 had n = 230 patients. The outcome variable is time (months) to progression, i.e., continued growth of the cancer. Several prognostic factors were identified as important predictors of progression. These include treatment, gender, patient's age, toxicity, and quality of life (QOL). Among these covariates, toxicity and QOL were missing for some patients. We model the covariate distribution of these missing covariates using a sequence of one-dimensional conditional distributions as in Ibrahim, Lipsitz, and Chen (1999), which we discuss in detail in Section

2.1. Our objective in the analysis of the LCCC 9719 data set was to select the most important predictors of SCLC progression and estimate the parameters of the best model. These selection and estimation processes can be done simultaneously by combining one of the two penalty functions, SCAD or ALASSO, with one of the two penalty estimates, these being the random effects penalty estimate or the IC_Q penalty estimate.

The rest of the article is organized as follows. Section 2 gives the general development of maximizing the penalized likelihood function and estimating penalty parameters. In Section 3, we characterize the asymptotic properties of the MPL estimator and IC_Q penalty selection procedures. Section 4 presents a simulation study that examines the finite sample performance of the MPL estimates and gives analyses of two lung cancer data sets. We conclude the article with some discussion in Section 5.

2. Variable Selection for the Cox Model with Missing Covariates

2.1 Model Formulation

Suppose that there are n independent observations $(T_1, c_1, \mathbf{z}_1, \mathbf{x}_1), \ldots, (T_n, c_n, \mathbf{z}_n, \mathbf{x}_n)$, where T_i is the time-to-the event, c_i is the censoring time, $(\mathbf{z}_i^T, \mathbf{x}_i^T)^T$ is a $p \times 1$ vector of covariates where $\mathbf{x}_i$ is a $(p - q) \times 1$ vector of fully observed covariates and $\mathbf{z}_i=(\mathbf{z}_{i,m}^T, \mathbf{z}_{i,o}^T)$ is a $q \times 1$ vector of partially observed covariates. Denote $y_i = \min(T_i, c_i)$, and let the vectors $\mathbf{z}_{i,o}$, and $\mathbf{z}_{i,m}$ denote the observed and missing components of $\mathbf{z}_i$, respectively. Let $\delta_i = 1\ \{T_i \leq c_i\}$ be the indicator of censoring, and $\mathcal{R}(t) = \{i: y_i \geq t\}$ be the set of subjects at risk at time t. Let $\mathbf{d}_{i,o} = (y_i, \delta_i, \mathbf{z}_{i,o}, \mathbf{x}_i)$ and $\mathbf{d}_{i,c} = (y_i, \delta_i, \mathbf{z}_i, \mathbf{x}_i)$ denote the observed data and complete data, respectively, for the ith observation. Throughout this article, we assume that the covariates are MAR, i.e., the probability of a missing covariate does not depend on any of the observed covariate values (Little and Rubin, 2002). We also assume that the parameters of the missing data mechanism are distinct from the sampling model, so that the missing data mechanism need not be modeled in the complete data likelihood. We specify the joint distribution of $(y_i, \delta_i, \mathbf{z}_i | \mathbf{x}_i)$ as a product of two conditional distributions,

$$f(y_i, \delta_i, \mathbf{z}_i|\mathbf{x}_i;\theta)=f(y_i, \delta_i|\mathbf{z}_i, \mathbf{x}_i;\theta)f(\mathbf{z}_i|\mathbf{x}_i;\theta),$$

where $\boldsymbol{\theta}$ includes all the unknown parameters. The generic label $f(u_1|u_2)$ is used to denote the conditional distribution of u_1 given u_2. We assume that the parameters of the distribution of the censoring times are distinct from those of the distribution of the survival times and that the distribution of the censoring times is independent of the unobserved covariates. Under these assumptions, the conditional distribution of (y_i, δ_i) given $(\mathbf{z}_i, \mathbf{x}_i)$ can be written as

$$f(y_i, \delta_i|\mathbf{z}_i, \mathbf{x}_i;\theta)=f_t(y_i|\mathbf{z}_i, \mathbf{x}_i;\theta)^{\delta_i} S_t(y_i|\mathbf{z}_i, \mathbf{x}_i;\theta)^{1-\delta_i} \times f_c(y_i|\mathbf{x}_i)^{1-\delta_i} S_c(y_i|\mathbf{x}_i)^{\delta_i},$$

where f_t, S_t, f_c, and S_c are the density and survival functions of the survival time and censoring time, respectively.

We assume a proportional hazards model (Cox, 1975) for the failure times, which assumes that the hazard of subject i at failure time y_i is $\lambda(y_i)\exp((\mathbf{z}_i^T, \mathbf{x}_i^T)\beta)$, where $\lambda(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. This allows the distribution of (y_i, δ_i) given $(\mathbf{z}_i, \mathbf{x}_i)$ to be written as

$$f(y_i, \delta_i|\mathbf{z}_i, \mathbf{x}_i;\beta, \Lambda) \propto \lambda(y_i)^{\delta_i}\exp\delta_i(\mathbf{z}_i^T, \mathbf{x}_i^T)\beta\exp\left\{-\Lambda(y_i)e^{(\mathbf{z}_i^T,\mathbf{x}_i^T)\beta}\right\}, \quad (1)$$

where $\Lambda(t)=\int_0^t \lambda(u)du$ is the cumulative baseline hazard function. Note that we have ignored all terms that are independent of $(\boldsymbol{\beta}, \Lambda)$ and $\mathbf{x}_i$. Finally, following Ibrahim et al. (1999), we write the distribution of $\mathbf{z}_i$ given $\mathbf{x}_i$ as

$$f(\mathbf{z}_i|\mathbf{x}_i;\alpha)=f(z_{i,q}|z_{i,(q-1)}, \ldots, z_{i,1}, \mathbf{x}_i;\alpha) \times \cdots \times f(z_{i,1}|\mathbf{x}_i;\alpha),$$

where $\boldsymbol{\alpha}$ are the parameters corresponding to the covariate distribution.

2.2 EM Algorithm for Maximizing the Penalized Likelihood

In the variable selection problem, our objective is to identify nonzero components of $\boldsymbol{\beta}$ in equation (1) and simultaneously estimate all other parameters while accounting for missing covariates. We propose to maximize the penalized likelihood function, given by

$$\ell(\theta) - n\sum_{j=1}^{p}\varphi_{\tau_j}(|\beta_j|)=\sum_{i=1}^{n}\ell_i(\theta) - n\sum_{j=1}^{p}\varphi_{\tau_j}(|\beta_j|), \quad (2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \Lambda)$, $\ell_i(\boldsymbol{\theta}) = \log \int f(y_i, \delta_i, \mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\theta})\, d\mathbf{z}_{i,m}$ is the observed-data log likelihood for the ith observation, τ_j is the penalty parameter corresponding to the jth regression coefficient, and the penalty function, $f_{\tau_j}(\cdot)$, is a nonnegative, nondecreasing, differentiable function on $(0, \infty)$ (Fan and Li, 2001; Zou, 2006). These properties ensure that the maximization of equation (2) results in certain estimates of $\boldsymbol{\beta}$ being zero (Antoniadis and Fan, 2001; Fan and Li, 2001). The regression coefficients that are estimated to be zero correspond to the covariates which are insignificant predictors of survival time, whereas other covariates are significant predictors.

Because the observed-data log-likelihood function usually involves intractable integration, we develop a Monte Carlo EM algorithm to compute the MPL estimator of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$, for each $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$. Let $\mathbf{D}_c$ and $\mathbf{D}_o$ denote the complete and observed data for all subjects, respectively, and let $L_c(\boldsymbol{\theta}|\mathbf{D}_c) = \log f(\mathbf{D}_c|\boldsymbol{\theta})$ denote the complete-data log-likelihood function. At the sth iteration, given $\boldsymbol{\theta}^{(s)}$, the E step is to evaluate the *penalized Q-function*

$$Q_\tau(\theta|\theta^{(s)})=Q(\theta|\theta^{(s)}) - n\sum_{j=1}^{p}\varphi_{\tau_j}(|\beta_j|), \quad (3)$$

where

$$\begin{aligned} Q(\theta|\theta^{(s)}) &= E\{L_c(\theta|\mathbf{D}_c)|\mathbf{D}_o, \theta^{(s)}\} \\ &= Q_1(\beta, \Lambda|\theta^{(s)}) + Q_2(\alpha|\theta^{(s)}), \end{aligned}$$

$$Q_1(\beta,\Lambda|\theta^{(s)})=\sum_{i=1}^{n}\int \log f(y_i,\delta_i|\mathbf{z}_i,\mathbf{x}_i;\beta,\Lambda)\times f(\mathbf{z}_{i,m}|\mathbf{d}_{i,o};\theta^{(s)})\,d\mathbf{z}_{i,m}, \quad \text{and} \tag{4}$$

$$Q_2(\alpha|\theta^{(s)})=\sum_{i=1}^{n}\int \log f(\mathbf{z}_i|\mathbf{x}_i;\alpha) f(\mathbf{z}_{i,m}|\mathbf{d}_{i,o};\theta^{(s)})\,d\mathbf{z}_{i,m}. \tag{5}$$

Because the integrals in equations (4) and (5) are often intractable, we approximate these integrals by taking a Markov chain Monte Carlo sample of size L from the density $f(\mathbf{z}_{i,m} \mid \mathbf{d}_{i,o}; \boldsymbol{\theta}^{(s)})$ (see Herring and Ibrahim, 2001). Let $\mathbf{z}_i^{(s,l)}=(\mathbf{z}_{i,m}^{(s,l)}, \mathbf{z}_{i,o})$, where $\mathbf{z}_{i,m}^{(s,l)}$ is the lth simulated value at the sth iteration of the algorithm. The integrals in equations (4) and (5) can be approximated as

$$\begin{aligned} Q_1(\beta,\Lambda|\theta^{(s)}) &\approx \sum_{i=1}^{n}\left[\delta_i\log\{\lambda(y_i)\}+\frac{1}{L}\sum_{l=1}^{L}\delta_i(\mathbf{z}_i^{(s,l)T},\mathbf{x}_i^T)\beta-\frac{1}{L}\sum_{l=1}^{L}\Lambda(y_i)\exp\{(\mathbf{z}_i^{(s,l)T},\mathbf{x}_i^T)\beta\}\right], \\ Q_2(\tau|\theta^{(s)}) &\approx \frac{1}{L}\sum_{i=1}^{n}\sum_{l=1}^{L}\log f(\mathbf{z}_i^{(s,l)}|\mathbf{x}_i;\alpha). \end{aligned} \tag{6}$$

The M-step involves maximizing $Q_\tau(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \Lambda)$. Rather than estimate the absolutely continuous function $\Lambda(t)$, $t \geq 0$, we estimate an increasing stepwise version of Λ. This involves maximizing with respect to $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and the parameters $\{\Lambda(x_i)$: $\delta_i = 1$ for $i = 1, \ldots n\}$. Using this parametrization, the maximizers of $Q_\tau(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$ are given by

$$\begin{aligned} \beta^{(s+1)} &= \operatorname*{argmax}_{\beta} \mathrm{PQ}_{1,\tau}(\beta|\theta^{(s)}), \\ \alpha^{(s+1)} &= \operatorname*{argmax}_{\alpha}\left\{\sum_{i=1}^{n}L^{-1}\sum_{l=1}^{L}\log f(\mathbf{z}_i^{(s,l)}|\mathbf{x}_i;\alpha)\right\}, \\ \lambda^{(s+1)}(y_i) &= \delta_i\left[\sum_{u\in\mathscr{R}(x_i)}\frac{1}{L}\sum_{l=1}^{L}\exp\{(\mathbf{z}_u^{(s,l)T},\mathbf{x}_u^T)\beta^{(s+1)}\}\right]^{-1}, \\ \Lambda^{(s+1)}(y_i) &= \sum_{u=1}^{n}\lambda(y_u)^{(s+1)}1\{y_u\leq y_i,\delta_u=1\}, \end{aligned}$$

where

$$\mathrm{PQ}_{1,\tau}(\beta|\theta^{(s)})=\mathrm{PQ}_1(\beta|\theta^{(s)})-n\sum_{j=1}^{p}\varphi_{\tau_j}(|\beta_j|), \quad \text{and}$$

$$\mathrm{PQ}_1(\beta|\theta^{(s)})=\sum_{i=1}^{n}\frac{1}{L}\sum_{l=1}^{L}\delta_i(\mathbf{z}_i^{(s,l)T},\mathbf{x}_i^T)1\beta-\sum_{i=1}^{n}\delta_i\log\left[\sum_{u\in\mathscr{R}(y_i)}\frac{1}{L}\sum_{l=1}^{L}\exp\{(\mathbf{z}_u^{(s,l)T}\mathbf{x}_u^T)\beta\}\right].$$

Maximizing $Q_2(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(s)})$ with respect to $\boldsymbol{\alpha}$ is straightforward and can be done using a standard optimization algorithm, such as the Newton–Raphson algorithm (Little and Schluchter, 1985; Schluchter and Jackson, 1989). Maximizing $\mathrm{PQ}_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$ with respect to $\boldsymbol{\beta}$, however, is very difficult because $\mathrm{PQ}_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$ is a nondifferentiable and nonconcave function of $\boldsymbol{\beta}$ (Zou and Li, 2008).

To maximize $PQ_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$, following Fan and Li (2001), a second-order Taylor's series approximation of $PQ_1(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$, centered at the value $\boldsymbol{\beta}^{(s)}$, is used. An expression for this approximation is given in Web Appendix A. Using this approximation, $PQ_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$ resembles a penalized weighted least squares regression, so algorithms for minimizing penalized least squares can be used. Such algorithms include the local quadratic approximation algorithm (Fan and Li, 2001) and the local linear approximation algorithm (Zou and Li, 2008). We use the local linear approximation algorithm because it reduces the computational cost of penalized maximizations (Zou and Li, 2008).

Using the approximation of $PQ_1(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$, let $\boldsymbol{\beta}^{(s+1)}$ be the maximizer of $PQ_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$. Because an approximation is used for $PQ_1(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$, $\boldsymbol{\beta}^{(s+1)}$ may not necessarily be the maximizer of $Q_{1,\tau}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(s)})$. Following the expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993), a value $\boldsymbol{\theta}^{(s+1)}$ can be produced, such that $Q_\tau(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)}) \geq Q_\tau(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^{(s)})$ rather than directly maximizing $Q_\tau(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. Therefore, we only need to obtain a $\boldsymbol{\beta}^{(s+1)}$ which satisfies $Q_{1,\tau}(\boldsymbol{\beta}^{(s+1)}|\boldsymbol{\theta}^{(s)}) \geq Q_{1,\tau}(\boldsymbol{\beta}^{(s)}|\boldsymbol{\theta}^{(s)})$. This process is iterated until convergence and the value at convergence is denoted as $\hat{\boldsymbol{\theta}}_\tau$. The value $\hat{\boldsymbol{\theta}}_\tau$ maximizes the penalized observed data log-likelihood function.

2.3 Penalty Parameter Selection Procedure

To ensure that $\hat{\boldsymbol{\theta}}_\tau$ has good properties, the penalty parameter $\boldsymbol{\tau}$ has to be appropriately selected. Two commonly used criteria for selection of the penalty parameter include the GCV and BIC criteria. These criteria cannot be easily computed in the presence of missing data, because they are functions of observed data quantities whose expressions require intractable integrals. Moreover, it has been shown in Wang et al. (2007) that even in the simple linear model, the GCV criterion can lead to significant overfit.

We propose two methods to select the penalty parameter: an IC_Q criterion and a random effects penalty estimation method. The IC_Q criterion (Ibrahim, Zhu, and Tang, 2008) selects the optimal $\boldsymbol{\tau}$ by minimizing

$$IC_Q(\boldsymbol{\tau}) = -2Q(\widehat{\theta_\tau}|\widehat{\theta_0}) + c_n(\widehat{\theta_\tau}),$$

where $\hat{\boldsymbol{\theta}}_0 = \text{argmax}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$ and $c_n(\boldsymbol{\theta})$ is a function of the data and the fitted model. For instance, if c_n equals twice the total number of parameters, then we obtain an AIC-type criterion; alternatively, we obtain a BIC-type criterion when $c_n(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta}) \times \log n$. Moreover, in the absence of missing data, $IC_Q(\boldsymbol{\tau})$ reduces to the usual AIC or BIC criteria. As in the EM algorithm, we can draw a set of samples from $f(\mathbf{z}_{i,m}|\mathbf{d}_{i,o}; \hat{\boldsymbol{\theta}}_0)$ for $i = 1, \ldots, n$ to to estimate $Q(\hat{\boldsymbol{\theta}}_\tau|\hat{\boldsymbol{\theta}}_0)$ for any $\boldsymbol{\tau}$.

The random effects penalty estimator is calculated under the assumption that the regression coefficients $\boldsymbol{\beta}$ are distributed as random effects in a hierarchical model. The parameter $\boldsymbol{\tau}$ can be regarded as a parameter in the distribution of $\boldsymbol{\beta}$, denoted by $f(\boldsymbol{\beta}|\boldsymbol{\tau}, n)$. Then, $\boldsymbol{\tau}$ can be estimated by maximizing the marginal likelihood with respect to $(\boldsymbol{\alpha}, \Lambda, \boldsymbol{\tau})$, which is given by

$$\int\prod_{i=1}^{n}\int f(y_i, \delta_i, \mathbf{z}_i|\mathbf{x}_i;\theta) f(\beta|\boldsymbol{\tau}, n)\, d\mathbf{z}_{i,m}\, d\beta = \prod_{i=1}^{n}\int f(\mathbf{d}_{i,o}|\theta) f(\beta|\boldsymbol{\tau}, n)\, d\beta, \tag{7}$$

where $f(\boldsymbol{\beta}|\boldsymbol{\tau}, n)$ is defined by

$$f(\beta|\boldsymbol{\tau}, n)=\prod_{j=1}^{p}\exp\{-n\varphi_{\tau_j}(|\beta_j|)\}/[C(\tau_j, n)],$$

and $C(\tau_j, n)$ is the normalizing constant of $f(\beta|\tau_j, n)$. The resulting estimate of $\boldsymbol{\tau}$, denoted by $\hat{\boldsymbol{\tau}}_{RE}$, from the maximization of equation (7), is the random effects penalty estimator. Treating the regression coefficients as missing data, the EM algorithm can be used to calculate $\hat{\boldsymbol{\tau}}_{RE}$.

We consider the SCAD and ALASSO penalty functions for estimating $\boldsymbol{\tau}_{RE}$. The ALASSO penalty is defined by

$$\varphi_{\tau_j}(|\beta_j|)=\tau_j|\beta_j| \text{ for } j=1,\ldots,p.$$

Typical values chosen for τ_j are $\tau_j = \tau_0|\hat{\beta}_j|^{-\gamma}$, where $\hat{\beta}_j$ is the unpenalized maximum likelihood (ML) estimate and $\gamma > 0$ is a pre-specified positive scalar. The SCAD penalty (Fan and Li, 2001) is a nonconcave function defined by $\varphi_\tau(0) = 0$ and for $|\beta| > 0$,

$$\varphi'_\tau(|\beta|)=\tau 1(|\beta| \le \tau)+\frac{(a\tau - |\beta|)_+}{a-1}1(|\beta|>\tau),$$

where t_+ denotes the positive part of t and $a = 3.7$. Because the integral of the negative exponential of the SCAD penalty is not finite, i.e., $\int_{-\infty}^{\infty}\exp\{-n\varphi_\tau(|\beta|)\}d\beta=\infty$, we use a truncated version of $p_\tau(|\beta|)$. This density is defined in Web Appendix A. For the ALASSO penalty, this truncation is not necessary because $\int_{-\infty}^{\infty}\exp\{-n\varphi_\tau(|\beta|)\}d\beta<\infty$. Because a closed form expression of $\hat{\boldsymbol{\tau}}_{RE}$ is unavailable for both the ALASSO and SCAD penalties, we use the Newton–Raphson algorithm along with the ECM algorithm to estimate $\hat{\boldsymbol{\tau}}_{RE}$.

Algorithms to estimate the IC_Q penalty estimate, the random effects penalty estimate, and the MPL estimate are given in Web Appendix A.

3. Theoretical Results

In this section, we establish the asymptotic theory of the MPL estimator and the consistency of the penalty estimation procedure based on IC_Q. Suppose $\beta=(\beta_{(1)}^T,\beta_{(2)}^T)^T$, where $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$ are, respectively, $p_1 \times 1$ and $p_2 \times 1$ subvectors such that $p = p_1 + p_2$. Let $\beta^*=(\beta_{(1)}^{*T},\beta_{(2)}^{*T})^T$ denote the true value of $\boldsymbol{\beta}$. Without loss of generality, we assume that $\beta_{(2)}^*=0$ and all of the components of $\beta_{(1)}^*$ are not equal to zero.

Let $\mathcal{S} = \{j_1, \ldots, j_d\}$ be a candidate model containing the j_1th, …, j_d th covariates. Thus, $\mathcal{S}_F = \{1, \ldots, p\}$ and $\mathcal{S}_T = \{1, \ldots, p_1\}$ denote the full and true covariate models, respectively. If $\mathcal{S}$ misses at least one important covariate, that is $\mathcal{S} \not\supset \mathcal{S}_T$, then $\mathcal{S}$ is referred to as an underfitted model; however, if $\mathcal{S} \supset \mathcal{S}_T$ and $\mathcal{S} \neq \mathcal{S}_T$, then $\mathcal{S}$ is an overfitted model. Suppose we only consider the selected covariates in $\mathcal{S}$. The unpenalized and penalized ML estimators of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \Lambda)^T$, denoted by $\hat{\boldsymbol{\theta}}_S$ and $\hat{\boldsymbol{\theta}}_\tau$, respectively, are defined as

$$\widehat{\theta}_s = \underset{\theta:\beta_j\neq 0,\forall j\in\mathcal{S}}{\operatorname{argmax}}\ \ell(\theta) \text{ and}$$
$$\widehat{\theta}_\tau = \underset{\theta}{\operatorname{argmax}} \left\{ \ell(\theta) - n\sum_{j=1}^{p} \varphi_{\tau_j}(|\beta_j|) \right\},$$

and particularly $\hat{\theta}_{\mathcal{S}_T} = \hat{\boldsymbol{\theta}}_0$. We obtain the following theorems whose assumptions and proofs can be found in Web Appendix A.

Theorem 1

Under Assumptions (C1)–(C7) in Web Appendix A, we have

- **a.** $\hat{\boldsymbol{\gamma}}_\tau - \boldsymbol{\gamma}^* = O_p(n^{-1/2})$ as $n \to \infty$, where $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ and $\boldsymbol{\gamma}^*$ is the true value of $\boldsymbol{\gamma}$;
- **b.** Sparsity: $P(\hat{\boldsymbol{\beta}}_{(2)\tau} = 0) \to 1$;
- **c.** Asymptotic normality: $\sqrt{n}\{(\widehat{\beta}_{(1)\tau}^T, \widehat{\alpha}_\tau^T)^T - (\beta_{(1)}^{*T}, \alpha^{*T})^T\}$ is asymptotically normal with mean and covariance matrix defined in Web Appendix A.

Theorem 1 states that by appropriately choosing the penalty $\boldsymbol{\tau}$, there exists a root-n estimator of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}_\tau$, and that this estimator must possess the sparsity property, i.e., $\hat{\boldsymbol{\beta}}_{(2)\tau} = 0$ in probability. Moreover, $(\widehat{\beta}_{(1)\tau}^T, \widehat{\alpha}_\tau^T)^T$ is asymptotically normal.

We investigate whether the $\mathrm{IC}_Q(\boldsymbol{\tau})$ criterion can consistently select the correct model. For each $\boldsymbol{\tau} \in R^{p+}$, $\hat{\boldsymbol{\beta}}_\tau$ naturally defines a candidate model $\mathcal{S}_\tau = \{j: \hat{\boldsymbol{\beta}}_{\tau,j} \neq 0\}$. Generally, $\mathcal{S}_\tau$ can be either underfitted, overfitted, or true. Therefore, R^{p+} can be partitioned into three mutually exclusive regions $R_u^{p+} = \{\boldsymbol{\tau} \in R^{p+}: \mathcal{S}_\tau \not\supset \mathcal{S}_T\}$, $R_t^{p+} = \{\boldsymbol{\tau} \in R^{p+}: \mathcal{S}_\tau = \mathcal{S}_T\}$, and $R_o^{p+} = \{\boldsymbol{\tau} \in R^{p+}: \mathcal{S}_\tau \supset \mathcal{S}_T, \mathcal{S}_\tau \neq \mathcal{S}_T\}$. Furthermore, if we can choose a reference penalty parameter sequence $\{\tau_n \in R^{p+}\}_{n=1}^{\infty}$, which satisfies the conditions of Theorem 1, then $\mathcal{S}_{\tau_n} = \mathcal{S}_T$ in probability.

To select $\boldsymbol{\tau}$, we first calculate

$$\begin{aligned} \mathrm{dIC}_Q(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) &= \mathrm{IC}_Q(\boldsymbol{\tau}_2) - \mathrm{IC}_Q(\boldsymbol{\tau}_1) \\ &= -2Q(\widehat{\theta}_{\tau_2}|\widehat{\theta}_0) + c_n(\widehat{\theta}_{\tau_2}) + 2Q(\widehat{\theta}_{\tau_1}|\widehat{\theta}_0) - c_n(\widehat{\theta}_{\tau_1}) \end{aligned}$$

for any two $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$. We assume $\mathcal{S}_{\tau_2} \supset \mathcal{S}_{\tau_1}$ and choose the model $\mathcal{S}_{\tau_1}$ resulting from using the penalty value $\boldsymbol{\tau}_1$ if $\mathrm{dIC}_Q(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) \geq 0$, otherwise we choose the model $\mathcal{S}_{\tau_2}$.

Define $\delta_Q(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = E\{Q(\theta^*_{\mathcal{S}_{\tau_1}}|\theta^*)\} - E\{Q(\theta^*_{\mathcal{S}_{\tau_2}}|\theta^*)\}$, and $\delta_c(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) = c_n(\hat{\boldsymbol{\theta}}_{\tau_2}) - c_n(\hat{\boldsymbol{\theta}}_{\tau_1})$, where $\theta^*_{\mathcal{S}}$ is defined in the supplementary document.

Theorem 2

Under Assumptions (C1)–(C7) in Web Appendix A, we have the following results.

- **a.** If for all $\mathcal{S}_\tau \not\supset \mathcal{S}_T$, $\liminf_n \delta_Q(\boldsymbol{\tau}, 0)/n > 0$ and $\delta_c(\boldsymbol{\tau}, 0) = o_p(n)$, then $\mathrm{dIC}_Q(\boldsymbol{\tau}, 0) > 0$ in probability.

b. If $E\{Q(\theta^*_{\mathscr{S}_{\tau_1}}|\widehat{\theta}_0)\} - E\{Q(\theta^*_{\mathscr{S}_{\tau_2}}|\widehat{\theta}_0)\} = O_p(n^{1/2})$ and $Q(\widehat{\theta}_{\tau_t}|\widehat{\theta}_0) - E\{Q(\theta^*_{\mathscr{S}_{\tau_t}}|\widehat{\theta}_0)\} = O_p(n^{1/2})$ for $t = 1, 2$, then $\mathrm{dIC}_Q(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) > 0$ in probability as $n^{-1/2}\delta_c(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1)$ converges to ∞ in probability.

c. If $Q(\hat{\boldsymbol{\theta}}_{\tau 1}|\hat{\boldsymbol{\theta}}_0) - Q(\hat{\boldsymbol{\theta}}_{\tau 2}|\hat{\boldsymbol{\theta}}_0) = O_p(1)$, then $\mathrm{dIC}_Q(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) > 0$ in probability as $\delta_c(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1)$ converges to ∞ in probability.

Theorem 2 has some important implications. Theorem 2(a) shows that $\mathrm{IC}_Q(\boldsymbol{\tau})$ chooses all significant covariates with probability 1. Because $\mathscr{S}_0 \subset R_t^p \cup R_o^p$, the optimal model selected by minimizing $\mathrm{IC}_Q(\boldsymbol{\tau})$ will not select a $\boldsymbol{\tau}$ with $\mathscr{S}_\tau \not\supset \mathscr{S}_T$ because $\mathrm{dIC}_Q(\boldsymbol{\tau}, 0) > 0$ in probability. Therefore, the $\mathrm{IC}_Q(\boldsymbol{\tau})$ criterion selects all significant covariates with probability tending to 1. Generally, the most commonly used $c_n(\boldsymbol{\theta})$, such as $2\dim(\boldsymbol{\theta})$, $\dim(\boldsymbol{\theta})\log(n)$, and $K \log\log(n)(K > 0)$, satisfy the condition $\delta_c(\boldsymbol{\tau}, 0) = o_p(n)$. The condition $\liminf_n n^{-1}\delta_Q(\boldsymbol{\tau}, 0) > 0$ ensures that $\mathrm{IC}_Q(\boldsymbol{\tau})$ chooses a model with large $E\{Q(\theta^*_{\mathscr{S}}|\theta^*)\}$. This condition is analogous to condition 2 in Wang et al. (2007), which elucidates the effect of underfitted models. The term $n^{-1}E\{Q(\theta^*|\theta^*)\} - n^{-1}E\{Q(\theta^*_{\mathscr{S}}|\theta^*)\}$ can be written as

$$n^{-1}\ell(\theta^*) - n^{-1}\ell(\theta^*_{\mathscr{S}}) + n^{-1}E\{H(\theta^*|\theta^*)\} - n^{-1}E\{H(\theta^*_{\mathscr{S}}|\theta^*)\}, \tag{8}$$

where

$$H(\theta_1|\theta_2) = \sum_{i=1}^{n} \int \log\{f(\mathbf{z}_{i,m}|\mathbf{d}_{i,o};\theta_1)\} f(\mathbf{z}_{i,m}|\mathbf{d}_{i,o};\theta_2)\, d\mathbf{z}_{i,m}. \tag{9}$$

By Jensen's inequality, the third and fourth terms of equation (8) are greater than zero and the first and second terms must be greater than zero for large n. Thus, $\liminf_n n^{-1}\delta_Q(\boldsymbol{\tau}, 0) \geq 0$ in probability.

If $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ have the same average $n^{-1}E\{Q(\theta^*_{\mathscr{S}_\tau}|\theta^*)\}$, that is, $\liminf_n n^{-1}\delta_Q(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) = 0$, then Theorem 2 (b) and (c) indicate that $\mathrm{IC}_Q(\boldsymbol{\tau})$ picks out the smaller model $\mathscr{S}_{\tau_1}$ when $\delta_c(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1)$ increases to ∞ at a certain rate (e.g., $\log(n)$). For example, for the BIC-type criterion, $\delta_c(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) = \{\dim(\theta_{\mathscr{S}_{\tau_2}}) - \dim(\theta_{\mathscr{S}_{\tau_1}})\}\log(n) \geq \log(n)$ because we assume $\mathscr{S}_{\tau_2} \supset \mathscr{S}_{\tau_1}$. The AIC-type criterion, for which $c_n(\boldsymbol{\theta}) = 2 \times \dim(\boldsymbol{\theta})$, however, does not satisfy this condition. Thus, similar to the AIC criterion with no missing data, $\mathrm{IC}_Q(\boldsymbol{\tau})$ with $c_n(\boldsymbol{\theta}) = 2 \times \dim(\boldsymbol{\theta})$ tends to overfit.

4. Numerical Studies

4.1 Example 1: Simulation Study

We demonstrate the performance of the MPL estimate using our proposed penalty estimators via simulations and compare them to the unpenalized ML estimator. Our objectives for these simulations were: (i) to compare the performance of the random effects and the IC_Q penalty estimators; (ii) to compare the performance of the SCAD and ALASSO penalty functions; and (iii) to determine how the comparisons in (i) and (ii) differ in the complete data and missing covariate settings.

We simulated data sets consisting of $n = 100$, 300, and 500 observations from the hazard model $\lambda(t|\mathbf{u}) = \lambda_0(t)\exp(\mathbf{u}^T\boldsymbol{\beta}^*)$, where $\lambda_0(t) = 1$, $\boldsymbol{\beta}^* = (0.8, 1, 0, 0, 0, 0, 0.6, 0)^T$, and the components of $\mathbf{u} = (u_1, \ldots, u_8)$ are standard normal and the correlation between u_i and u_j is $\rho^{|i-j|}$ with $\rho = 0.5$. The censoring times, T_i, were selected to have an exponential

distribution with mean $v\exp(0.6u_7)$, where v is uniformly distributed over [4, 6]. Under these conditions, each simulated data set has about 30% of its survival times right censored. For the data sets with missing data, the missing covariates $\mathbf{z}_i = (u_{i1}, u_{i2})^T$ were taken to be MAR and $\mathbf{x}_i = (u_{i3}, u_{i4}, u_{i5}, u_{i6}, u_{i7}, u_{i8})^T$ were completely observed. The covariate distribution for the missing covariates is $\mathbf{z}_i \sim N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$, the $\mathbf{z}_i$'s are independent where $\mu_i=(\mu_{1i}, \mu_{2i}), \mu_{si}=\alpha_{s0}+\sum_{j=3}^{8}\alpha_{sj}u_{is}$, for $s = 1, 2$ and $\boldsymbol{\Sigma}$ is an unstructured 2×2 covariance matrix. The missing data mechanism was given by $f(r_{i1}, r_{i2} |\mathbf{x}_i, \boldsymbol{\xi}) = f(r_{i1}|r_{i2}, \mathbf{x}_i, \boldsymbol{\xi}_1)\, f(r_{i2}|\mathbf{x}_i, \boldsymbol{\xi}_2)$, where

$$f(r_{i1}=1|\mathbf{x}_i, \boldsymbol{\xi}_1)=\frac{\exp(\boldsymbol{\phi}_{i1})}{1+\exp(\boldsymbol{\phi}_{i1})},$$
$$\boldsymbol{\phi}_{i1}=\xi_{10}+\sum_{j=1}^{5}\xi_{1j}x_{ij}+\xi_{16}y_i, \quad \text{and}$$

$$f(r_{i2}=1|r_{i1}, \mathbf{x}_i, \boldsymbol{\xi}_2)=\frac{\exp(\boldsymbol{\phi}_{i2})}{1+\exp(\boldsymbol{\phi}_{i2})},$$
$$\boldsymbol{\phi}_{i2}=\xi_{20}+\sum_{j=1}^{5}\xi_{2j}x_{ij}+\xi_{26}y_i+\xi_{27}r_{i1}.$$

The values $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ were selected to achieve 25% missingness.

For each simulated data set, the unpenalized ML and the MPL estimates using the SCAD and ALASSO penalties were computed using the random effects and IC_Q estimators. For the IC_Q estimates, the BIC-type criterion, $c_n(\boldsymbol{\theta}) = \dim(\boldsymbol{\theta}) \times \log n$, was used. For the simulated data sets with no missing data, the IC_Q criterion is equivalent to BIC. To compute the penalty estimates and MPL estimate, 1000 Monte Carlo iterations were used within each iteration of EM. Initially, simulations with Monte Carlo samples of 5000 and 10,000 iterations were computed but the resulting estimates did not differ with those using 1000 Monte Carlo iterations, and therefore, samples of 1000 iterations were used to lessen the computational demand of the MPL procedure. Different starting values were selected to ensure that the algorithm converges to the global maximum. For the ALASSO penalty, we set $\tau_j = \tau_0|\hat{\beta}_{j0}|^{-1}$, where $\hat{\beta}_{j0}$ is the unpenalized ML estimate, while for the SCAD penalty we let $\tau_j = \tau_0$, for all j, where in both cases τ_0 was selected using the random effects and IC_Q penalty selection methods. For each estimate, $\hat{\boldsymbol{\beta}}_\tau$, the mean squared error $(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta})E(\mathbf{uu}^T)(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta})$ was computed. The ratio of the model error of an MPL estimate to that of the unpenalized ML estimate was calculated. The median of these ratios over the 100 simulations is reported in the column MRME. Also, the average mean square error (MME) across all simulations was computed. Additionally, the average number of coefficients correctly estimated to be zero and incorrectly estimated to be zero are reported in the columns "Correct" and "Incorrect," respectively. All of these statistics were also calculated for the true model, which is denoted as "True."

The results of the simulations are presented in Table 1. The SCAD and ALASSO penalty estimators using the random effects and IC_Q penalty selection methods are denoted as SCAD-RE, SCAD-IC_Q, ALASSO-RE, and ALASSO-IC_Q, respectively. The results indicate that when the sample size is small ($n = 100$), the SCAD-RE and ALASSO-IC_Q have the smallest model error. When the sample size is relatively large ($n = 300$ and $n = 500$), the ALASSO-IC_Q estimator has the smallest model error. For all the estimators, as the sample size gets larger, the model error gets smaller but it decreases at a slower rate than that of the unpenalized ML estimate. Some of the estimators had smaller error than that of the true

model. A possible explanation for this is that because the number of parameters increases as the sample size gets larger, the difference in the number of parameters between any model and the true model is relatively small in value. Comparatively, the ALASSO estimators tended to have larger overfit compared to the SCAD estimators. The IC_Q estimators had larger overfit compared to the random effects penalty estimators for the SCAD penalty but not for the ALASSO penalty. In particular, the ALASSO-RE estimator showed significant overfit.

4.2 Example 2: Veterans Administration Lung Cancer Data

We applied the proposed methodology to the well known Veterans Administration lung cancer data set of Kalbfleisch and Prentice (2002). Although these data have no missing covariates, we analyzed these data to compare complete data results with scenarios based on hypothetical missing data. The result of this analysis is available in Web Appendix A.

4.3 Example 3: Small Lung Cancer Data

We revisit the lung cancer data discussed in Section 1. As mentioned in Section 1, the covariates x_{i1} = treatment (2 arms: A and B, coded as 1 and 0), x_{i2} = gender (female and male, coded as 0 and 1), and x_{i3} = age in years were fully observed for all patients, and z_{i1} = highest grade toxicity (recorded by cycle) (2 levels: 0 versus > 0, coded as 0 and 1), and z_{i2} = quality of life (QOL) score were missing. The missing data fraction for toxicity and QOL individually were 28.2% and 35.2%, respectively, with 52.7% of the data containing missing information on at least one of these covariates. We assume that z_{i1} and z_{i2} are MAR and consider the two covariate distributions used in Chen, Ibrahim, and Shao (2009). The first distribution, called model 1, is specified by assuming

$[z_{i1}, z_{i2}|\mathbf{x}_i]=[z_{i1}|z_{i2}, \mathbf{x}_i][z_{i2}|\mathbf{x}_i]$, $[z_{i1}|z_{i2}, \mathbf{x}_i] \sim \text{Bernoulli}(1, \frac{\exp(\eta_i)}{1+\exp(\eta_i)})$, and $[z_{i2}|\mathbf{x}_i] \sim N(\boldsymbol{\mu}_i, \sigma^2)$ for $i = 1, \ldots, n$, where $\eta_i=\eta_0+\sum_{j=1}^{3}\eta_j x_{ij}+\eta_4 z_{i2}$, and $\mu_i=\mu_0+\sum_{j=1}^{3}\mu_s x_{ij}$. For the second covariate distribution, called model 2, we assume $[z_{i1}, z_{i1}|\mathbf{x}_i] = [z_{i2}|z_{i1}, \mathbf{x}_i][z_{i1}|\mathbf{x}_i]$, in which we specify a normal linear regression model for $[z_{i2}|z_{i1}, \mathbf{x}_i]$ and a logistic regression model for $[z_{i1}|\mathbf{x}_i]$.

The results of the analyses are presented in Table 2. For both models 1 and 2, the ALASSO-RE estimator identifies treatment and toxicity as significant predictors of survival of SCLC. The results of these estimators are consistent with the results of the unpenalized ML analysis where toxicity and treatment are the only covariates that are strongly significant (p-value ≤ 0.01). The SCAD and ALASSO-IC_Q estimators did not identify any covariates as significant predictors.

5. Discussion

We have proposed a general method to simultaneously perform model selection and estimation in the Cox regression model with MAR covariates. Under some regularity conditions and appropriate rates of the penalty parameter, we have shown that the MPL estimate possesses sparsity and asymptotic normality properties. We have developed two methods to select the penalty parameter, the IC_Q penalty estimator, and the random effects penalty estimator. Under an appropriate choice of $c_n(\cdot)$, we have shown that the IC_Q penalty estimate can choose all significant predictors with probability 1.

Simulation results have shown that the SCAD penalty function with the random effects penalty estimator performs well when the noise level is low, whereas it performs poorly when the noise level is high. Overall, the SCAD penalty performs better when it is used with the random effects penalty estimator whereas the ALASSO performed better when it is used with the IC_Q criterion. The ALASSO penalty shows significant overfit in the small sample

simulations and this overfit is also present in the real data analyses. In the presence of missing data, there seems to be significant underfit compared to the analysis with no missing data. The differences in the results between the penalty functions and penalty selection methods indicate that sensitivity analyses should be performed between the IC_Q and random effects penalty estimates and between the SCAD and ALASSO penalty functions.

A disadvantage of penalized likelihood methods is that they do not provide a measure of model uncertainty, i.e., the probability of selecting each model in the model space. Other methods, such as Bayesian model averaging (Hoeting et al., 1999), or other Bayesian methods in general, provide estimates of posterior model probabilities. However, implementation of fully Bayesian methods can be difficult in many cases, because it requires specifying priors for the parameters in the covariate distribution of all the models in the model space as well as calculating marginal likelihoods and enumerating all of the models in the model space. Alternatively, unlike MPL methods, Bayesian methods do not give an estimate of the parameters of the "best" model. An MPL estimate, however, is equal to the posterior mode of a fully Bayesian analysis with the prior $f(\beta, \alpha, \Lambda) \propto \prod_{j=1}^{p} \exp\{-n \sum_{j=1}^{p} \varphi_{\tau_j}(|\beta_j|)\}$. Therefore, the algorithm proposed in Section 2.2 to maximize the penalized likelihood can be easily modified to obtain the posterior mode in a fully Bayesian analysis.

The method proposed in this article only considers the $p < n$ setting, therefore generalizations of our method to the $p > n$ and $p \gg n$ settings need to be studied. Although we have only applied our method to data sets with dozens of covariates, we believe that it can be applied to data sets with hundreds of covariates with any type of missingness because our method is very similar to the algorithm used in Ibrahim et al. (1999). As p and n get large, however, certain computational issues can arise which make the implementation of our method difficult. For instance, it is important make sure that the EM algorithm converges to the global maximum of the penalized likelihood function. This can be ensured by starting the algorithm from multiple starting values. When p and n are large, it is easier to use the ALASSO penalty function along with the random effects penalty estimate because a closed form expression for the conditional maximizer of the penalty parameter is available. This allows easy implementation of the ECM algorithm to estimate the penalty parameter.

Many other aspects of this work warrant further research and investigation. As it stands, calculating the penalty estimate is computationally demanding. The random effects penalty estimate is easier to compute than the IC_Q penalty estimate. The theoretical properties of the random effects penalty estimate, however, need to be investigated, whereas the theoretical properties of the IC_Q penalty estimate are established. Alternative methods which select the penalty parameter based on optimizing some easy-to-compute criterion such as DIC (Celeux et al., 2006) or a modification of IC_Q can be investigated. For example, one could select the penalty parameter which minimizes $-2Q(\tilde{\boldsymbol{\theta}}_\tau|\hat{\boldsymbol{\theta}}_0) + c_n(\tilde{\boldsymbol{\theta}}_\tau)$, where $\tilde{\theta}_\tau = \text{argsup}_\theta\{Q(\theta|\widehat{\theta}_0) - n\sum_{j=1}^{p} \varphi_{\tau_j}(|\beta_j|)\}$. This method is less computationally intensive because $\tilde{\boldsymbol{\theta}}_\tau$ does not require as many iterations to compute compared to the IC_Q penalty estimate. We will formally study these issues in future work.

Acknowledgments

The authors thank the editor, the associate editor, and two referees for several suggestions and editorial changes, which have greatly improved the article. This research was partially supported by NIH grants GM 70335, CA 74015, and by United States Environmental Protection Agency—National Center for Computational Toxicology through the Curriculum in Toxicology—University of North Carolina, under Cooperative Training Program CR83323710. This work was supported in part by NSF grants SES-06-43663 and BCS-08-26844 and NIH grants UL1-RR025747-01 and R21AG033387 to Dr Zhu.

References

Antoniadis Z, Fan J. Regularization of wavelet approximations. Journal of the American Statistical Association. 2001; 96:939–955.

Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Analysis. 2006; 4:651–674.

Chen HY. Double-semiparametric method for missing covariates in Cox regression models. Journal of the American Statistical Association. 2002; 97:565–576.

Chen HY, Little RJA. Proportional hazard hegression with missing covariates. Journal of the American Statistical Association. 1999; 94:896–908.

Chen MH, Ibrahim JG, Shao QM. Model identifiability for the Cox regression model with applications to missing covariates. Journal of Multivariate Analysis. 2009 in press.

Cox DR. Partial likelihood. Biometrika. 1975; 362:269–276.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Annals of Statistics. 2002; 30:74–99.

Herring AH, Ibrahim JG. Likelihood-based methods for missing covariates in the Cox proportional hazards model. Journal of the American Statistical Association. 2001; 96:292–302.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science. 1999; 14:382–417.

Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. Journal of the Royal Statistical Society, Series B. 1999; 61:173–190.

Ibrahim JG, Zhu H, Tang N. Model selection criteria for missing data problems via the EM algorithm. Journal of the American Statistical Association. 2008; 103:1648–1658. [PubMed: 19693282]

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. New York: John Wiley; 2002.

Lipsitz SR, Ibrahim JG. Estimation with correlated censored survival data with missing covariates. Biostatistics. 2000; 1:315–327. [PubMed: 12933512]

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. New York: John Wiley; 2002.

Little RJA, Schluchter M. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika. 1985; 72:497–512.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80:267–278.

Paik MC, Tsai W. On using Cox proportional hazards model with missing covariates. Biometrika. 1997; 84:579–593.

Pettitt AN, Tran TT, Haynes MA, Hay JL. A Bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants. Journal of the Royal Statistical Society, Series A. 2006; 169:97–144.

Schluchter M, Jackson K. Log-linear analysis of censored survival data with partially observed covariates. Journal of the American Statistical Association. 1989; 84:42–52.

Socinski MA, Schell MJ, Peterman A, Bakri K, Yates S, Gitten R, Unger P, Lee J, Lee J, Tynan M, Moore M, Kies M. Phase III trial comparing duration of therapy versus continuous therapy followed by second-line therapy in advanced-stage IIIB/IV non-small-cell lung cancer. Journal of Clinical Oncology. 2002; 20:1335–1343. [PubMed: 11870177]

Wang H, Leng C. Unified LASSO estimation by least squares approximation. Journal of the American Statistical Association. 2007; 102:1039–1048.

Wang H, Li R, Tsai CL. Tuning parameter selector for the smoothly clipped absolute deviation method. Biometrika. 2007; 94:553–568. [PubMed: 19343105]

Zhang H, Lu W. Adaptive Lasso for Cox's proportional hazard model. Biometrika. 2007; 94:691–703.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in noncancave penalized likelihood models. Annals of Statistics. 2008; 36:1509–1533. [PubMed: 19823597]

Table 1

Simulation results for Cox regression model with no missing covariates and covariates MAR comparing SCAD and ALASSO penalty functions with random effects and IC_Q penalty estimates. No miss is the estimate from data with no missing covariates. MAR is the estimate from data with covariates missing at random. MRME is the median relative model error, MME is the median model error, and n is simulated data sample size.

		No miss (MAR)			
				# of 0 coefficients	
Model	**Method**	**MRME**	**MME**	**Correct**	**Incorrect**
n = 100	MLE	–	0.179 (0.240)	–	–
	SCAD-RE	0.373 (0.513)	0.055 (0.113)	4.73 (4.78)	0.04 (0.07)
	SCAD-IC_Q	0.441 (0.581)	0.060 (0.127)	4.80 (4.72)	0.05 (0.04)
	ALASSO-RE	0.447 (0.570)	0.076 (0.147)	3.66 (3.69)	0.01 (0.02)
	ALASSO-IC_Q	0.464 (0.628)	0.069 (0.145)	4.66 (4.60)	0.03 (0.04)
	True	0.334 (0.420)	0.043 (0.098)	5.00 (5.00)	0.00 (0.00)
n = 300	MLE	–	0.060 (0.068)	–	–
	SCAD-RE	0.574 (0.510)	0.031 (0.029)	4.77 (4.92)	0.00 (0.00)
	SCAD-IC_Q	0.602 (0.515)	0.034 (0.031)	4.91 (4.94)	0.00 (0.00)
	ALASSO-RE	0.615 (0.684)	0.036 (0.046)	3.60 (3.53)	0.00 (0.00)
	ALASSO-IC_Q	0.573 (0.693)	0.031 (0.039)	4.79 (4.80)	0.00 (0.00)
	True	0.574 (0.531)	0.031 (0.030)	5.00 (5.00)	0.00 (0.00)
n = 500	MLE	–	0.039 (0.040)	–	–
	SCAD-RE	0.577 (0.563)	0.022 (0.023)	4.82 (4.90)	0.00 (0.00)
	SCAD-IC_Q	0.581 (0.571)	0.022 (0.023)	4.95 (4.00)	0.00 (0.00)
	ALASSO-RE	0.631 (0.758)	0.026 (0.034)	3.75 (3.64)	0.00 (0.00)
	ALASSO-IC_Q	0.525 (0.815)	0.018 (0.031)	4.94 (4.77)	0.00 (0.00)
	True	0.574 (0.575)	0.021 (0.024)	5.00 (5.00)	0.00 (0.00)

Table 2

Maximum penalized likelihood estimates of LCCC 9719 small lung cancer data of models 1 and 2

	Model 1[a] (Model 2[b])				
	SCAD		ALASSO		
Variable	**RE**	**IC_Q**	**RE**	**IC_Q**	**MLE[c]**
Treatment	0.000 (0.000)	0.000 (0.000)	0.357 (0.316)	0.000 (0.000)	$0.472^{**}(0.472^{**})$
Gender	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.188 (0.186)
Age	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	−0.026 (−0.025)
Toxicity	0.000 (0.000)	0.000 (0.000)	0.438 (0.244)	0.000 (0.000)	$1.025^{**}(1.027^{**})$
QOL	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	−0.050 (−0.055)

[a] Is estimate from model 1

[b] Is estimate from model 2

** indicates p-value < 0.01